

# Supplementary Material of AttentionNet: Aggregating Weak Directions for Accurate Object Detection

Donggeun Yoo      Sunggyun Park      Joon-Young Lee\*      Anthony S. Paek      In So Kweon  
dgyoo@rcv.kaist.ac.kr      sunggyun@kaist.ac.kr      jylee@rcv.kaist.ac.kr      apaek@lunit.io      iskweon@kaist.ac.kr  
KAIST      KAIST      KAIST      Lunit Inc.      KAIST

This supplementary material contains extra experiments, example images, and details of AttentionNet [2].

## 1. Effectiveness of the Second Augmentation Rule

The second rule in the augmentation for training regions is as follows.

A positive region can include multiple instances, but a target instance must occupy the biggest area.  
Within a cropped region, an area of the target instance is at least 1.5-times larger than that of the other instances.

This rule is important to separate multiple instances that are overlapped. Table 1 and Fig. 1 show the impact of this rule.

Table 1: Average precisions (%) with/without the second augmentation rule on PASCAL VOC 2007 “person”.

Method	The second rule	AP(%)
AttentionNet	No.	51.3
AttentionNet	Yes.	<b>61.7</b>
AttentionNet + Refine	No.	52.7
AttentionNet + Refine	Yes.	<b>65.0</b>

---

\*This work was done when he was in KAIST. He is currently working in Adobe Research.

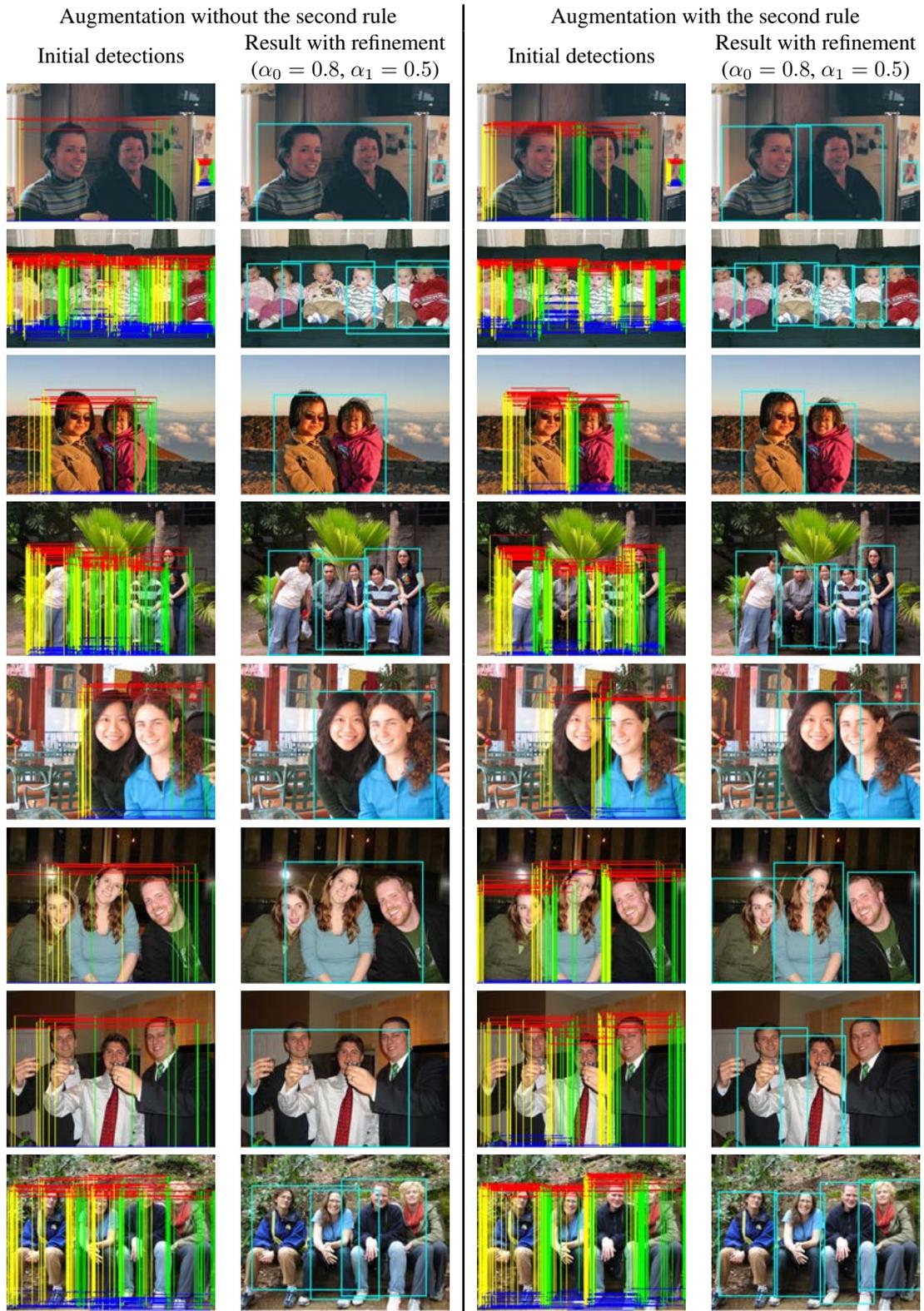


Figure 1: Detection examples with/without the second augmentation rule on PASCAL VOC 2007 “person”.

## 2. Performance to Number of Scales

Fig. 2 shows the performance to the number of scales. An image of a scale (e.g. 2) has two times larger resolution than that of the previous scale (e.g. 1). As shown in this figure, more than 6 scales are enough to achieve the best performance.

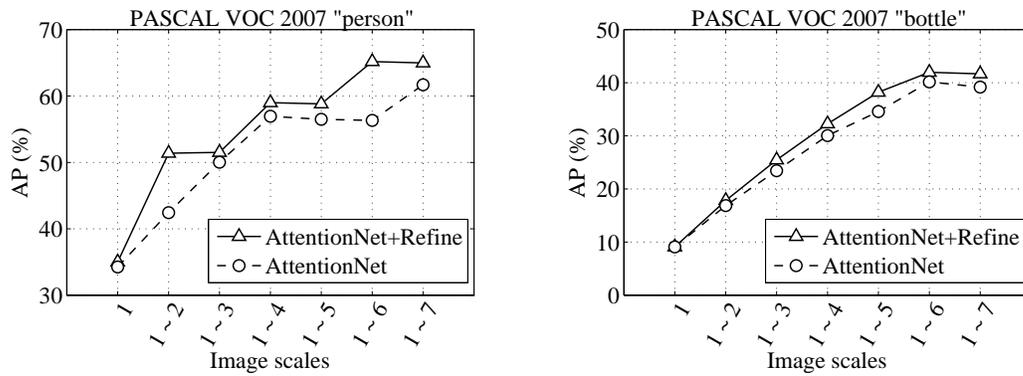


Figure 2: Performance to the number of scales on PASCAL VOC 2007 “person” (left) and “bottle” (right). The tick label of the horizontal axis is the combination of scales.

### 3. Details of Merging Initial Detections.

When we merge the initially detected bounding boxes as drawn in Fig. 3-(a), we reject the isolated bounding boxes which are not merged with other bounding boxes, because they are prone to be outliers from our multi-scale sliding window scheme.

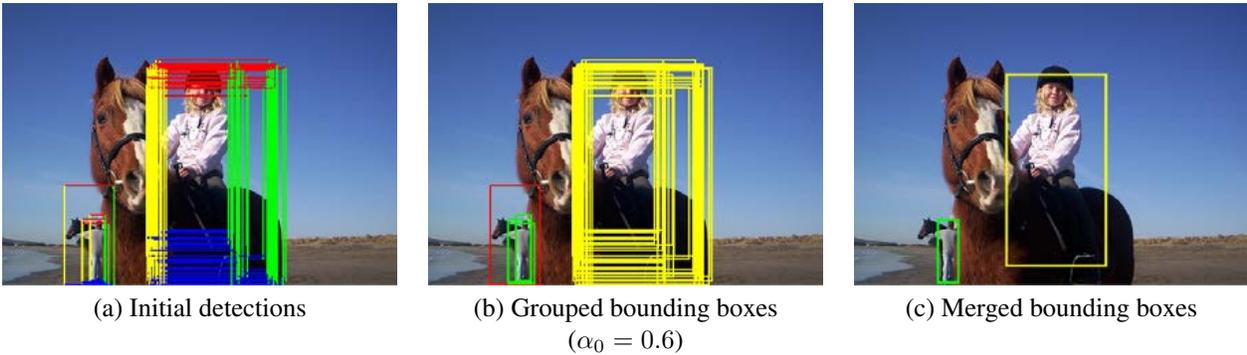


Figure 3: A real example of the initial merge procedure. A red bounding box in (b) is an outlier to be rejected, because it is not grouped with other bounding boxes.  $\alpha_0$  is a value of intersection over union (IoU) for the initial merge.

## 4. Qualitative Comparison

We show our detection examples and comparisons. We compare our detection results with Region-CNN (R-CNN) results, obtained from the source code<sup>1</sup> provided by Girshick *et al.* [1]. *We draw all bounding boxes detected by AttentionNet without any score threshold*, while only those that achieved greater than 30% AP in R-CNN are drawn. Through Fig. 4 to Fig. 12, we show both cases when our results are either superior or inferior to R-CNN.

---

<sup>1</sup><https://github.com/rbgirshick/rcnn>

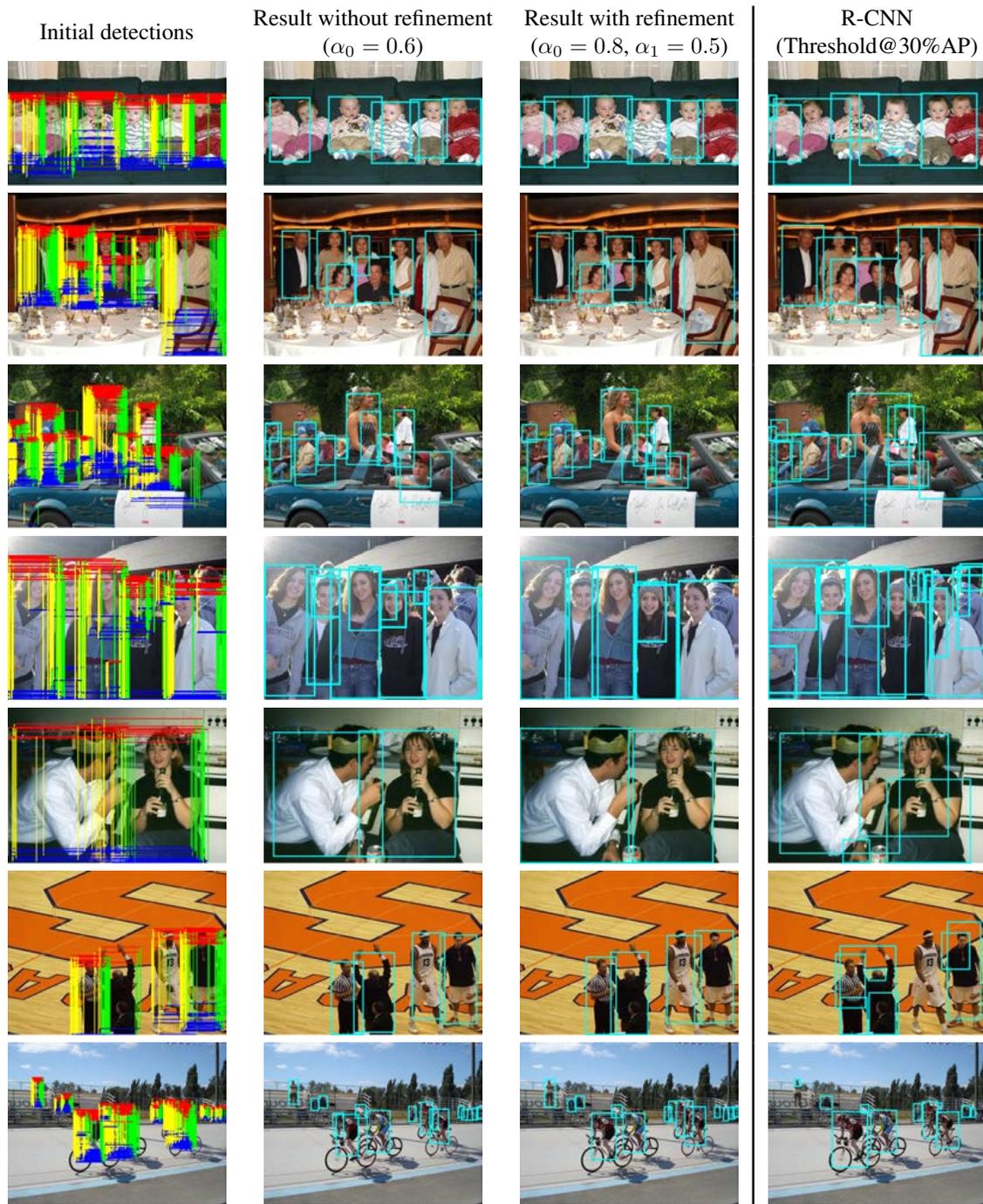


Figure 4: Examples when our result is **superior** to that of R-CNN in PASCAL VOC 2007 “person”.

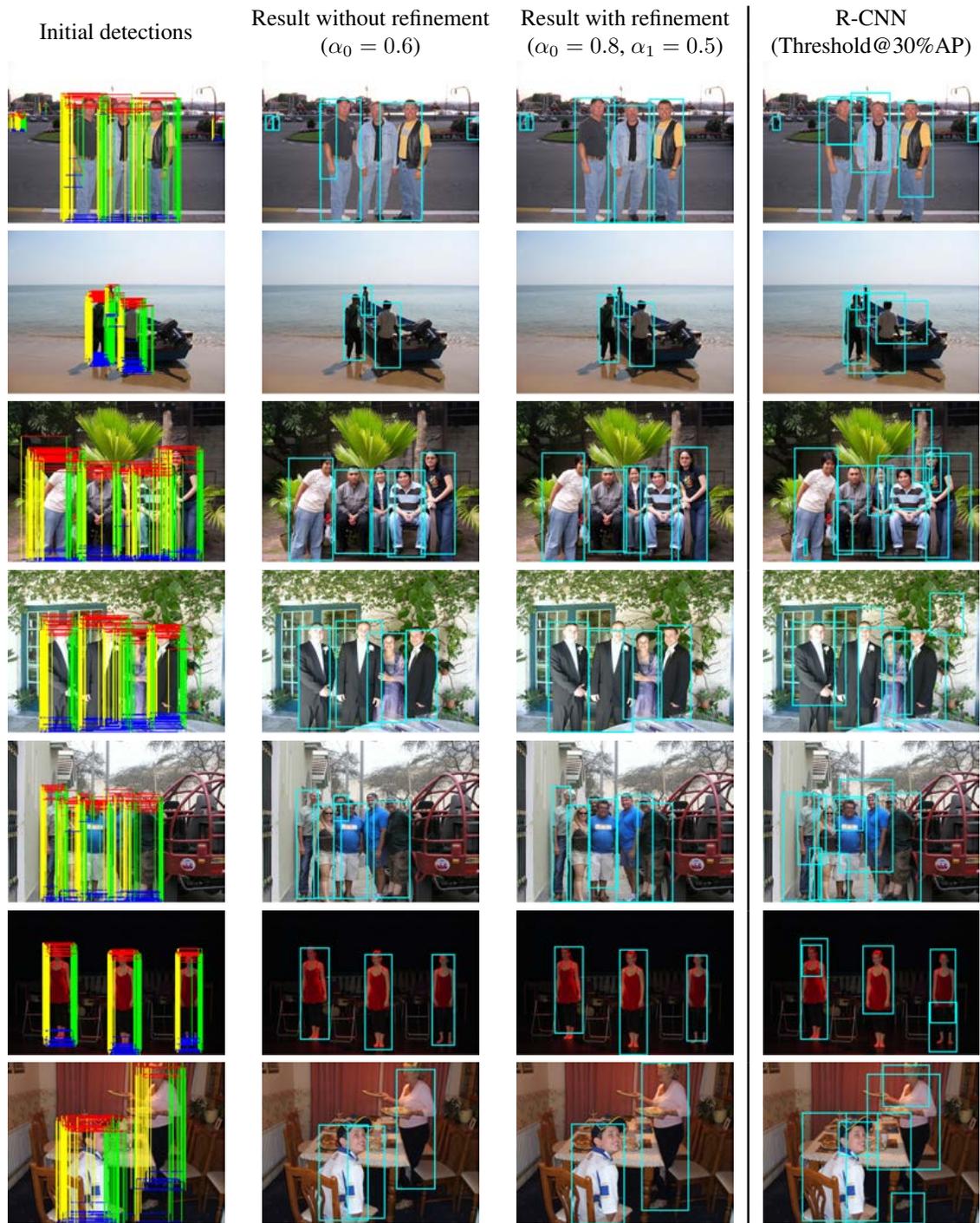


Figure 5: Examples when our result is **superior** to that of R-CNN in PASCAL VOC 2007 “person”.

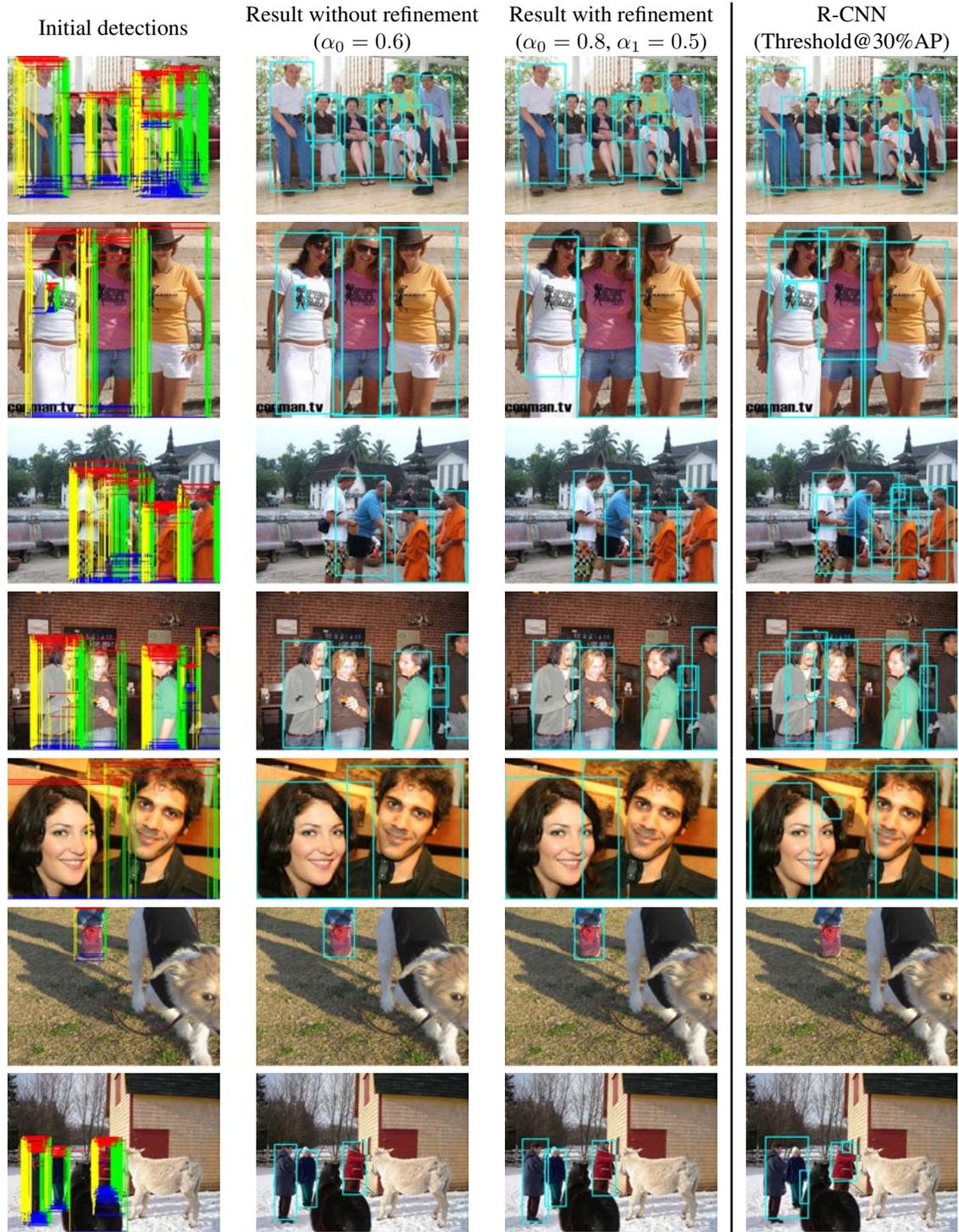


Figure 6: Examples when our result is **superior** to that of R-CNN in PASCAL VOC 2007 “person”.

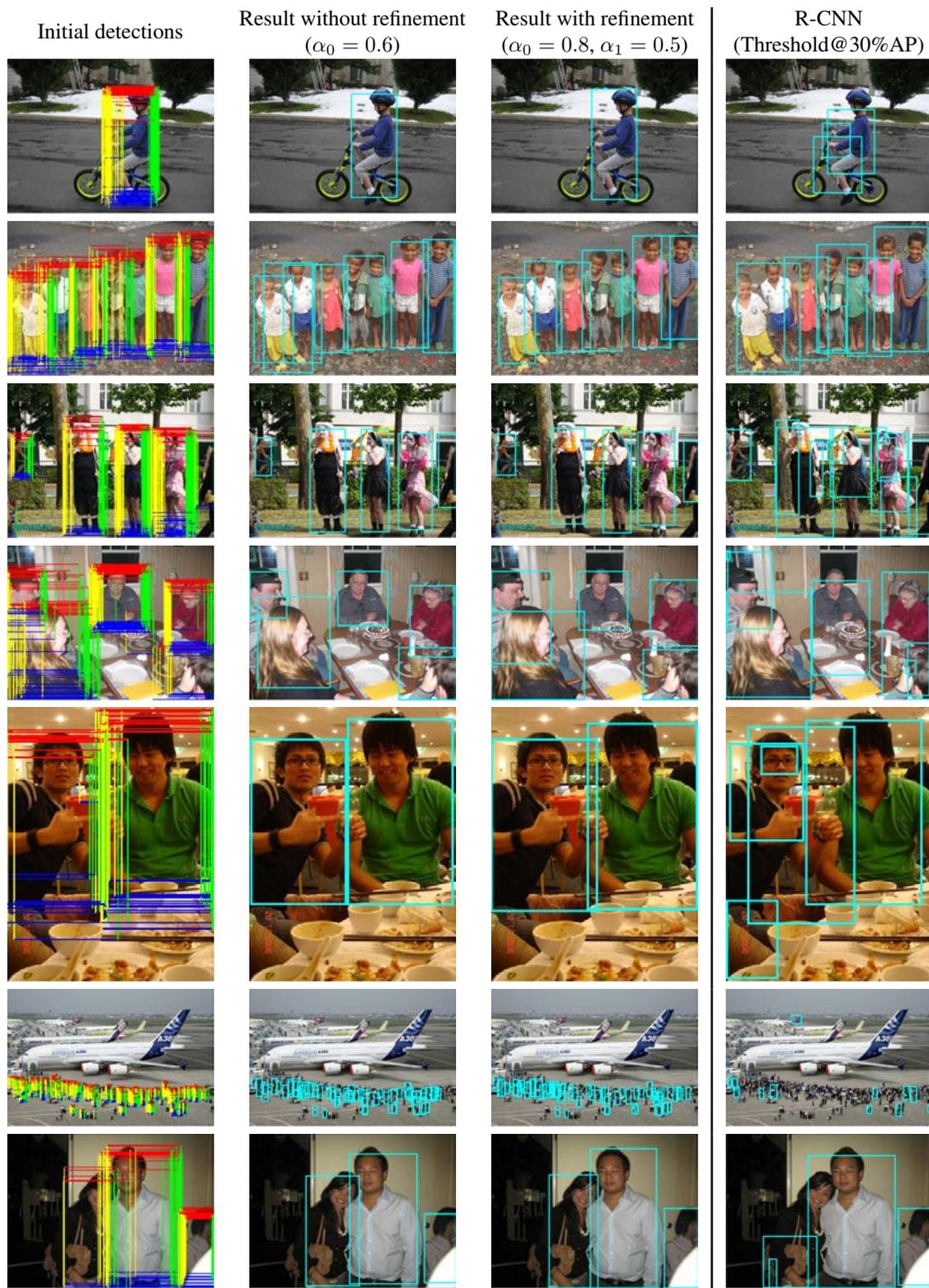


Figure 7: Examples when our result is **superior** to that of R-CNN in PASCAL VOC 2007 “person”.

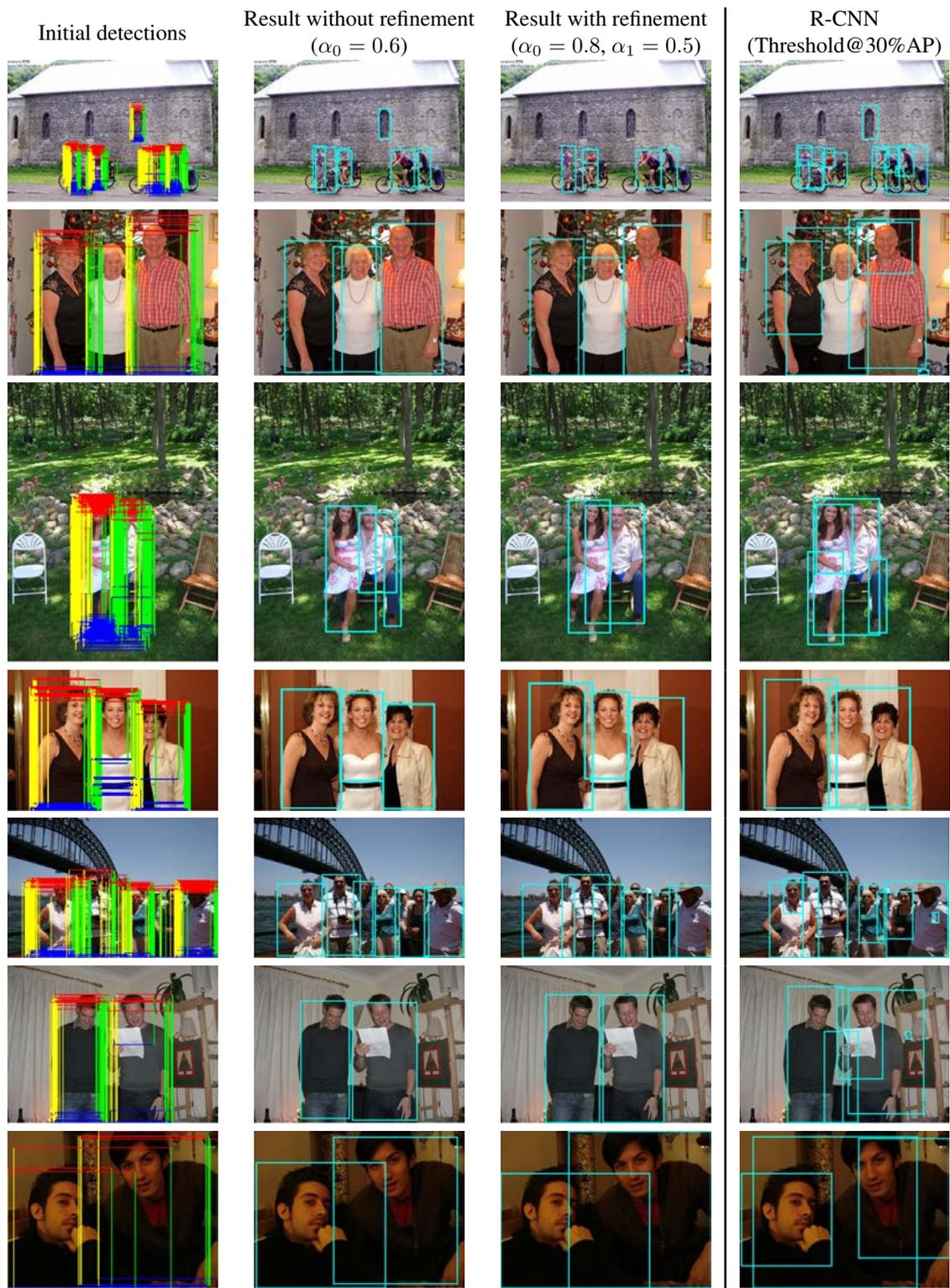


Figure 8: Examples when our result is **superior** to that of R-CNN in PASCAL VOC 2007 “person”.

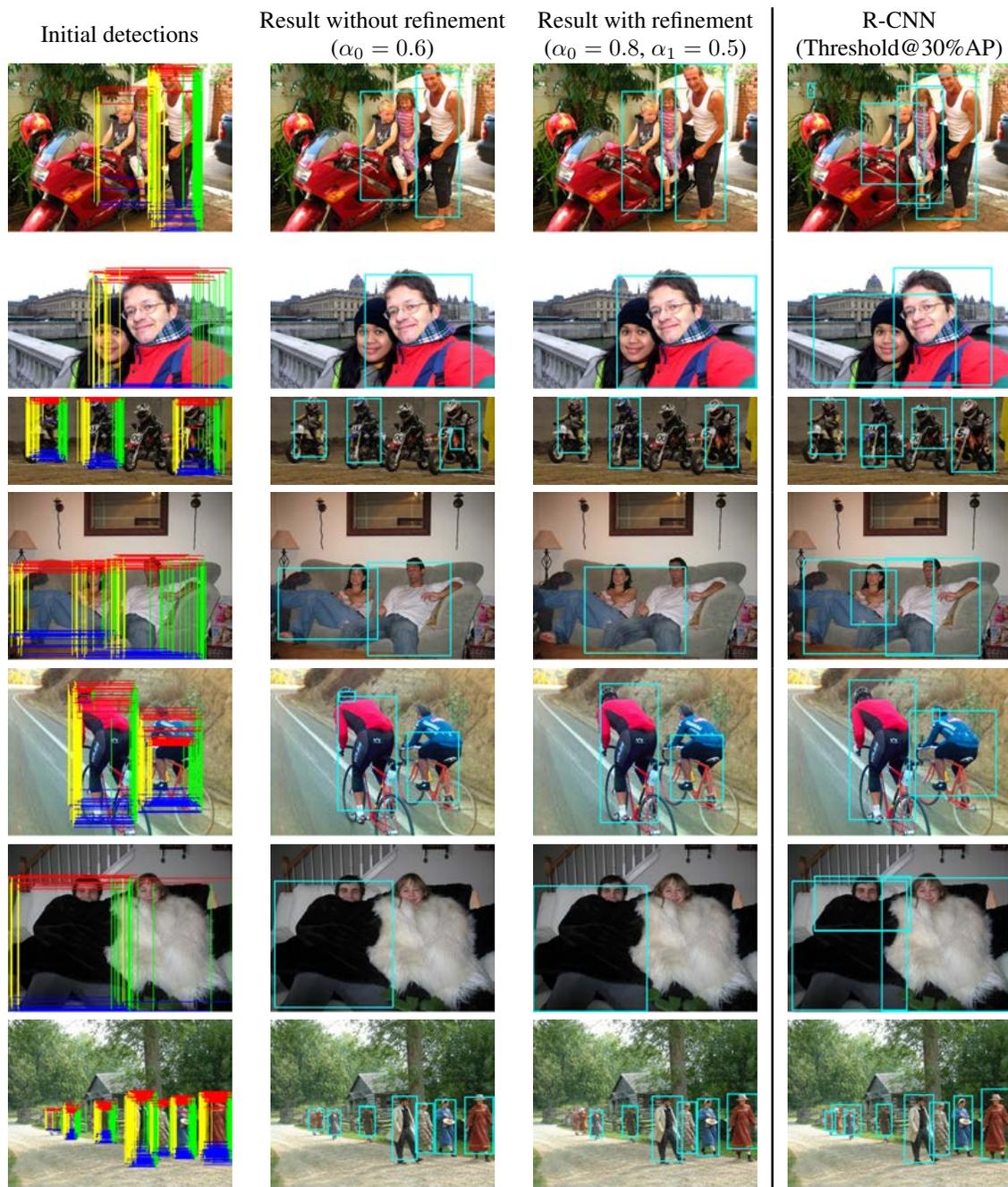


Figure 9: Examples when our result is **inferior** to that of R-CNN in PASCAL VOC 2007 “person”.

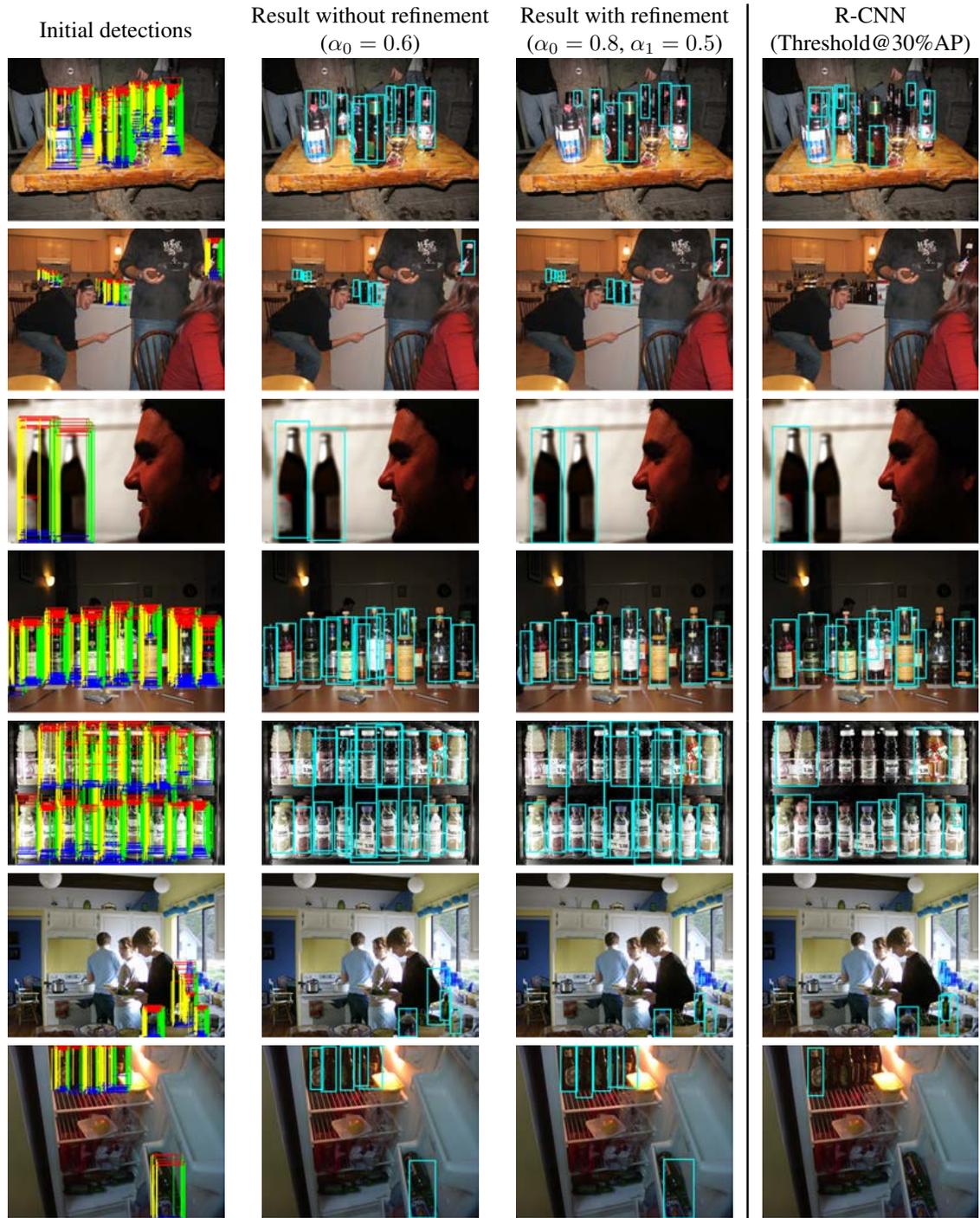


Figure 10: Examples when our result is **superior** to that of R-CNN in PASCAL VOC 2007 “bottle”.

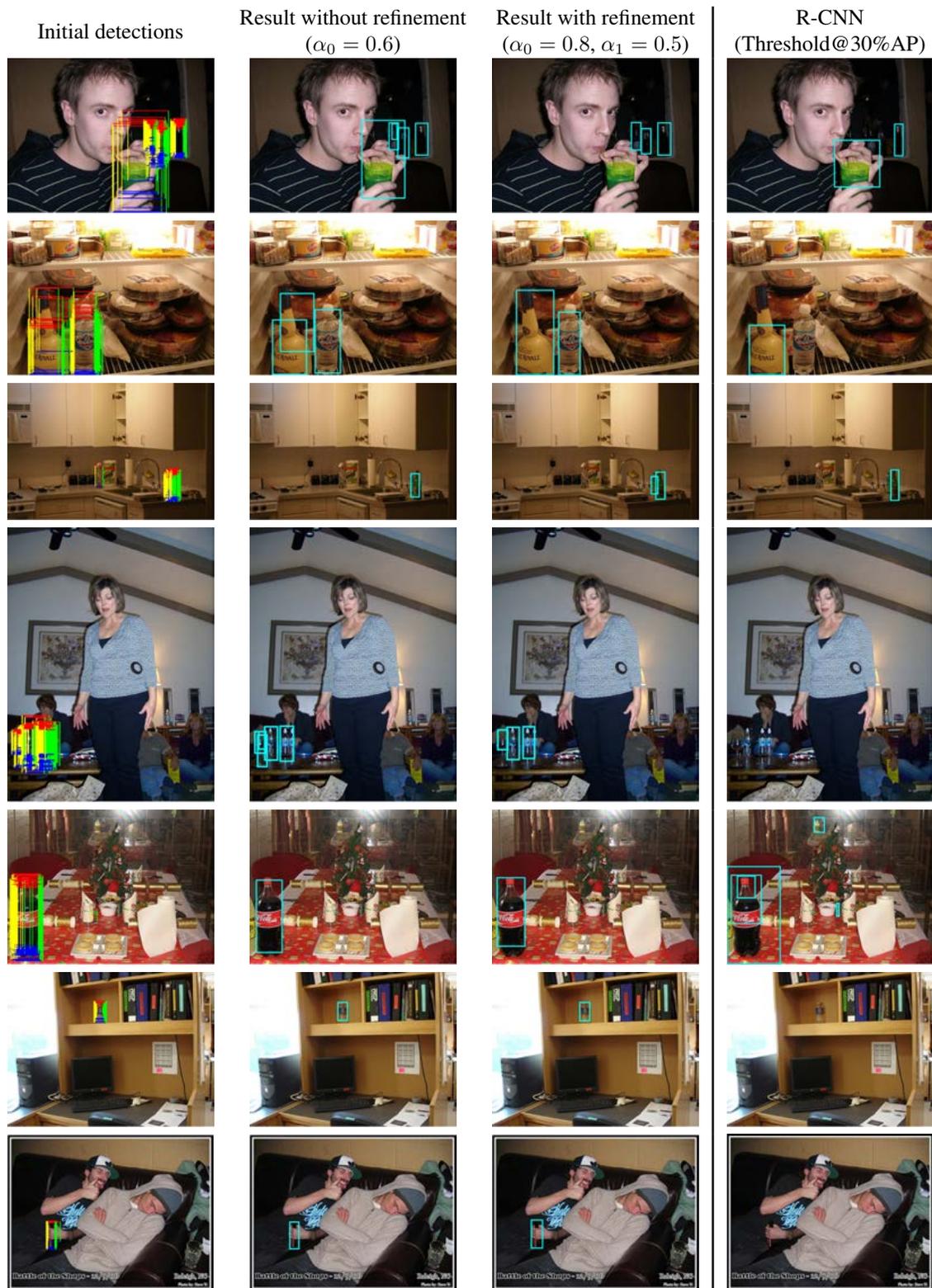


Figure 11: Examples when our result is **superior** to that of R-CNN in PASCAL VOC 2007 “bottle”.

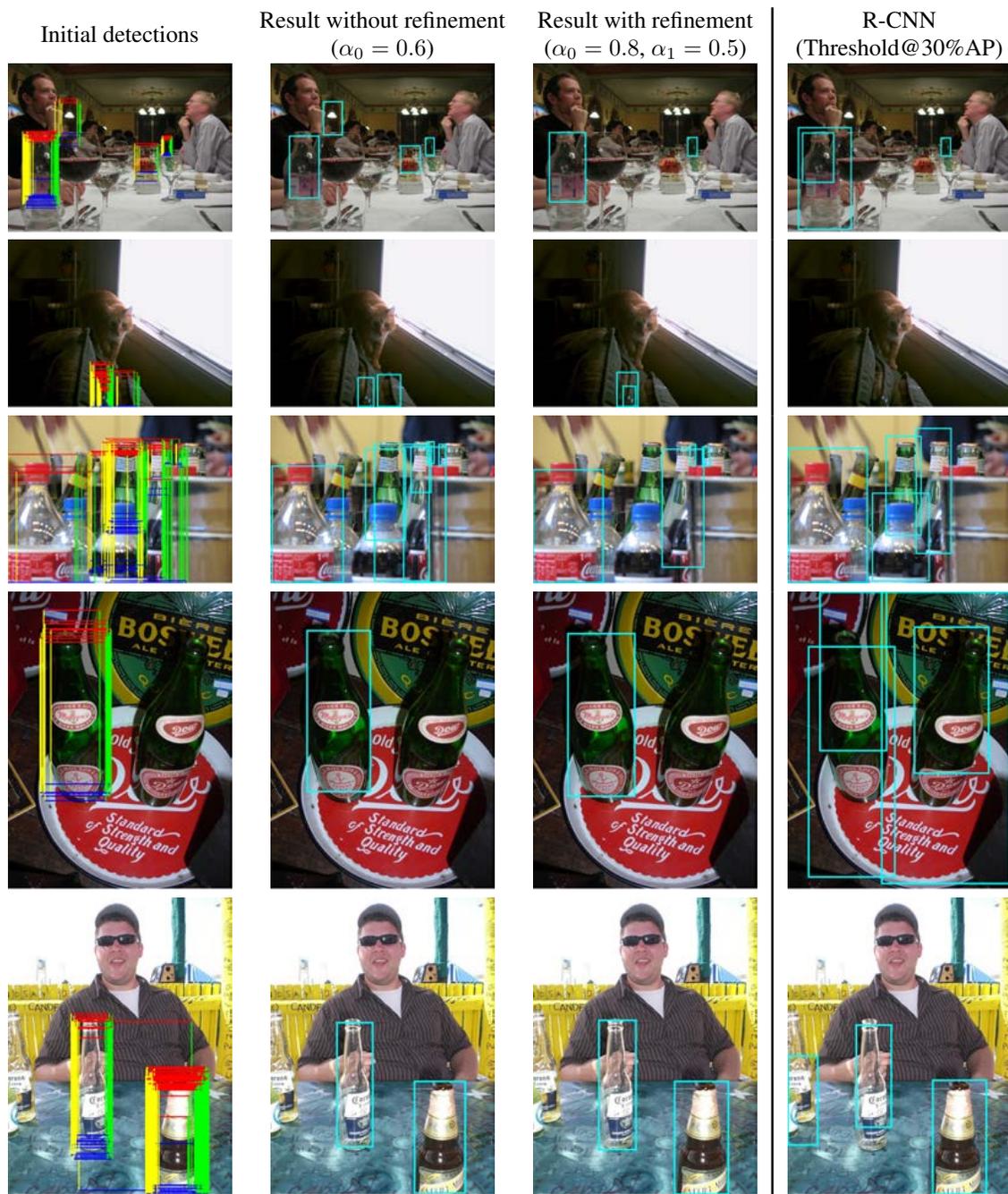


Figure 12: Examples when our result is **inferior** to that of R-CNN in PASCAL VOC 2007 “bottle”.

## References

- [1] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.
- [2] D. Yoo, S. Park, J.-Y. Lee, A. S. Paek, and I. S. Kweon. Attentionnet: Aggregating weak directions for accurate object detection. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2015.